Revealing Codes

June 8, 2004
By David A. Karp

What do your Microsoft Office documents say about you? SCO Group, a Utah-based business software company, found out the hard way when it recently filed suit against DaimlerChrysler and AutoZone. The suit's text had been created in Microsoft Word with Track Changes enabled, which users commonly do so they can see what edits were applied and by whom. Unfortunately, as is too often the case, no one had removed the tracking—which revealed, among other things, that Bank of America had been the original intended defendant.

Every Microsoft Excel, Word, and PowerPoint document contains a variety of information that remains present but hidden until you remove it or someone else extracts it. And therein lies the problem: If you plan on sharing or publishing your Office document, you may be sharing more than you intend.

The list of hidden data in an Office document may include everything from your name and e-mail address to deleted text, revision marking, and even the locations of related files on your computer. While some of the data is needed for collaboration features to work, other bits and pieces are not so vital. And whether or not these things are stored in your files depends largely (but not entirely) on settings in your Office applications.

For instance, you can change a setting to remove the information (commonly called metadata) displayed in the document summary. Select Options from the Tools menu, choose the Security tab, and check the "Remove personal information from this file on save" option.

Office apps will keep track of the revisions you (and others) make to your documents if you select Track Changes from the Tools menu, but you'll be able to see them only if you also choose Markup from the View menu. As the file is modified, any text that is deleted will continue to be stored in the file, along with the name of the person who deleted it. It's easy to overlook that deleted text if you choose to view the final (without markup) version of the document, so don't forget to banish any deleted text permanently. You can get rid of specific selections of deleted text by right-clicking the text and selecting "Accept Deletion," or remove them all at once by clicking Accept All Changes on the Reviewing toolbar.

If you really want to rid your documents of sensitive information, use one of the following tools.

Microsoft's Remove Hidden Data add-in (rhdtool.exe) is a command-line utility and macro available at office.microsoft .com. This tool will automatically remove all of the reviewing data and personal information, as well as about 30 types of hidden data that you probably didn't even know were stored in your files (routing slips, e-mail headers, and file paths to embedded objects, to name a few).

To use the add-in from within Word, you'll need to first save your file and then select Remove Hidden Data from the File menu. Or, use the command-line utility—the syntax is described in the included offrhdreadme.htm file—to remove hidden data from multiple files at once. Keep in mind that the add-in is a little buggy, and requires patience to get past the stalls and quirks of this Microsoft freebie.

If you create Web pages with Office, you'll want to look at Office 2000 HTML Filter 2.0 (office.microsoft.com). Not only will this tool remove nearly all hidden data from your Office-generated HTML pages, it will purge the extraneous HTML tags that clutter your Web content. It will also significantly shrink your files, decreasing their size and download time by as much as 50 percent.

If you grasp the implications of situations like the one SCO faced, you'll want something more automated than Microsoft's free tools. Both Esquire Innovations iScrub (www.esqinc.com) and Workshare Protect (www.workshare.net) include add-ins for Microsoft Outlook that ensure that all Office documents sent as e-mail attachments are cleansed of potentially sensitive information.

But no automated solution is perfect, mostly because of the ease with which these tools can be circumvented. Conversely, such a system might remove collaboration data from a file on its way to an editor. (Tip: put that document in a ZIP file to make it immune to tampering.) Likewise, an e-mail-based tool won't prevent someone from posting an uncleansed Word file on a Web site.